# Simple or Complex? Learning to Predict Readability of Bengali Texts

**Susmoy Chakraborty**[1*], **Mir Tafseer Nayeem**[1*], **Wasi Uddin Ahmad**[2]
[1]Ahsanullah University of Science and Technology
[2]University of California, Los Angeles
susmoyaust36@gmail.com, mir.nayeem@alumni.uleth.ca, wasiahmad@ucla.edu

## Abstract

Determining the readability of a text is the first step to its simplification. In this paper, we present a readability analysis tool capable of analyzing text written in the Bengali language to provide in-depth information on its readability and complexity. Despite being the $7^{th}$ most spoken language in the world with *230* million native speakers, Bengali suffers from a lack of fundamental resources for natural language processing. Readability related research of the Bengali language so far can be considered to be narrow and sometimes faulty due to the lack of resources. Therefore, we correctly adopt document-level readability formulas traditionally used for U.S. based education system to the Bengali language with a proper age-to-age comparison. Due to the unavailability of large-scale human-annotated corpora, we further divide the document-level task into sentence-level and experiment with neural architectures, which will serve as a baseline for the future works of Bengali readability prediction. During the process, we present several human-annotated corpora and dictionaries such as a document-level dataset comprising *618* documents with 12 different grade levels, a large-scale sentence-level dataset comprising more than *96K* sentences with simple and complex labels, a consonant conjunct count algorithm and a corpus of *341* words to validate the effectiveness of the algorithm, a list of *3,396* easy words, and an updated pronunciation dictionary with more than *67K* words. These resources can be useful for several other tasks of this low-resource language.[1]

## Introduction

The term *"Readability"* measures how much energy the reader will have to expend in order to understand a writing at optimal speed and find interesting. Readability measuring formulas, such as Automated Readability Index (ARI) (Senter and Smith 1967), Flesch Reading Ease (Flesch 1948), and Dale–Chall Formula (Dale and Chall 1948; Chall and Dale 1995) calculate a score that estimates the grade level or years of education of a reader based on the U.S. education system, which is illustrated in Figure 1. These formulas are still used in many widely known commercial readability measuring

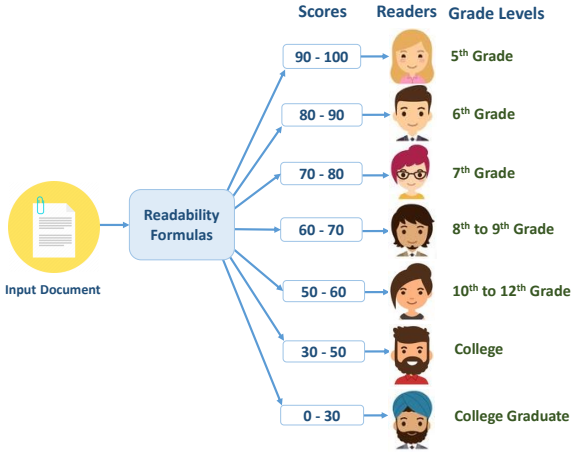


Figure 1: Readability prediction task.

tools such as *Grammarly*[2] and *Readable*.[3] This measurement plays a significant role in many places, such as education, health care, and government (Grigonyte et al. 2014; Rets et al. 2020; Meng et al. 2020). Government organizations use it to ensure that the official texts meet a minimum readability requirement. For instance, the Department of Insurance at Texas has a requirement that all insurance policy documents have a Flesch Reading Ease (Flesch 1948) score of 40 or higher, which translates to the reading level of a first-year undergraduate student based on the U.S. education system.[4] A legal document which is hard to read can lead someone to sign a contract without understanding what they are agreeing to. Another common usage area is the healthcare sector to ensure the proper readability of the care and treatment documents (Grigonyte et al. 2014). Better readability will attract visitors or readers of different websites or blogs, whereas poor readability may decrease the number of readers (Meng et al. 2020). Readability measures are also often used to assess the financial documents such as annual reports of a company's economic performance so that the information is more transparent to the reader (Loughran and McDonald

[*]Equal contribution. Listed by alphabetical order.

[1]We make our code & dataset publicly available at https://github.com/tafseer-nayeem/BengaliReadability for reproduciblity.

[2]https://www.grammarly.com/blog/readability-scores/
[3]https://readable.com/
[4]https://www.tdi.texas.gov/pubs/pc/pccpfaq.html

2014). *Dyslexia* is a disorder that causes difficulties with skills associated with learning, namely reading and writing, which affects up to 20% of the general population. Readability formulas have been applied to measure the difficulty of reading texts for people with dyslexia (Fourney et al. 2018).

The scores from readability formulas have been generally found to correlate highly with the actual readability of a text written in the English language. The adaptation of readability formulas to no-English texts is not straightforward. Measuring readability is also essential for every non-English language, but not all of the readability formulas mentioned above are language-independent. These formulas require some resources like a 3000-word list, which is easily understandable by fourth-grade American students, syllable counting dictionary, stemmer, lemmatizer etc. Resource availability for Natural Language Processing (NLP) research is an obstacle for some low-resource-languages (e.g., Bengali). In this paper, we aim to develop a readability analysis tool for the Bengali Language. Bengali is the native language of Bangladesh, also used in India (e.g., West Bengal, Tripura) and has approximately 230 million native speakers.[5] Despite being the $7^{th}$ most spoken language in the world, Bengali suffers from a lack of fundamental resources for NLP (Chowdhury et al. 2021). For a low resource language like Bengali, the research in this area so far can be considered to be narrow and sometimes incorrect. Islam, Mehler, and Rahman (2012); Sinha et al. (2012) tried to adapt the formula-based approaches used for the English language. Unfortunately, it isn't straightforward as these formulas are developed for U.S. based education system[6] and which predicts U.S. grade level of the reader. Since the Bangladeshi education system and grade levels[7] are different from U.S., therefore, the mapping is faulty and led to incorrect results. There is a strong relationship between reading skills and human cognition, which varies depending on different age groups (Riddle 2007). Therefore, to eliminate this incompatibility, in this paper, we map grade level to different age groups to present age-to-age comparison. Moreover, (Sinha and Basu 2016; Islam, Rahman, and Mehler 2014; Sinha et al. 2012) used traditional machine learning models to address this task on a very small scale dataset, which isn't publicly available. There are readability analysis tools available for *English* (Napolitano, Sheehan, and Mundkowsky 2015), *Arabic* (Al-Twairesh et al. 2016), *Italian* (Okinina, Frey, and Weiss 2020), and *Japanese* (Sato, Matsuyoshi, and Kondoh 2008) language. Unfortunately, no such tool is available for Bengali language that can validate the readability of a text. On the other hand, there is no large-scale human annotated readability analysis dataset available to train supervised neural models for this extremely low-resource language. Our main contributions are summarized as follows:

- To the best of our knowledge, we first design a comprehensive system for Bengali readability analysis, which includes datasets, human-annotated corpora and dictionaries, an algorithm, models, and a tool capable of proving in-depth readability information of the texts written in the Bengali language (see Figure 4).
- We correctly adopt document-level readability formulas traditionally used for U.S. based education system to the Bengali education system with a proper age-to-age comparison. We present a document level dataset consisting of 618 documents with 12 different grade levels to evaluate the adaptation of these formulas to the Bengali language (see Table 1).
- Due to the long-range dependencies of RNNs and the unavailability of large-scale human-annotated corpora, we further divide the document-level task into sentence-level and present a large-scale dataset consisting of 96,335 sentences with simple and complex labels to experiment with supervised neural models (see Table 2). We design neural architectures and make use of all available pre-trained language models of Bengali Language.
- We also propose an efficient algorithm for counting consonant conjuncts form a given word. We present a human-annotated corpus comprising 341 words with varying difficulties to evaluate the effectiveness of this algorithm.
- We design a readability analysis tool capable of analyzing text written in the Bengali language to provide in-depth information on its readability and complexity which would be useful for educators, content writers or editors, researchers, and readers (see Figure 4).

## Related Works

### English and Other Languages

There has been a lot of work on readability analysis for the English language, some of which are: Automated Readability Index (Senter and Smith 1967), Flesch Reading Ease (Flesch 1948), Flesch–Kincaid Grade Level (Kincaid 1975), Gunning Fog Index (Kincaid 1975), Dale–Chall Formula (Dale and Chall 1948; Chall and Dale 1995), and SMOG Grade (Mc Laughlin 1969). To calculate the readability score from these formulas, we need to extract various types of information from the input text, e.g., average sentence length, average word length, number of characters, number of syllables, and number of difficult words. The primary takeaway from the common formulas is simple. Using shorter sentences with shorter and more common words will typically make a piece of writing easier to understand. Other than using these traditional formulas, there have been a lot of recent work on readability analysis of English language (Vajjala and Lučić 2018; Dueppen et al. 2019; Rets et al. 2020; Meng et al. 2020), but especially in recent years, the scope of readability analysis research has been broadened towards other languages such as *Russian* (Reynolds 2016), *Japanese* (Sato 2014), *French* (Seretan 2012), *Swedish* (Grigonyte et al. 2014), *Polish* (Broda et al. 2014), *Arabic* (El-Haj and Rayson 2016), *Vietnamese* (Nguyen and Uitdenbogerd 2019), and *German* (Battisti et al. 2020).

[5] https://w.wiki/57

[6] https://w.wiki/Zoy

[7] https://www.scholaro.com/pro/Countries/bangladesh/Education-System

## Bengali Language

**Formula Based Approaches** The research in this area for Bengali Language is narrow and sometimes faulty. Das and Roychoudhury (2006) created a miniature readability model with one and two parametric fits using multiple regression for Bengali text readability analysis. They used only seven paragraphs from seven documents. They used two parameters, such as average sentence length and number of syllables per 100 words, which is responsible for representing text as easy or difficult. Sinha et al. (2012); Sinha and Basu (2016) proposed two readability formulas for each Bengali and Hindi text using regression analysis, which are similar to readability formulas used for the English language. Readability formulas were also applied to Bangladeshi textbooks by Islam, Mehler, and Rahman (2012). They extracted three types of features from data such as lexical features, entropy-based features, and Kullback-Leibler Divergence-based features. Unfortunately, the scores returned from these readability formulas approximate what grade level of U.S. based education system someone will need in order to be able to read a piece of text easily. Since the Bangladeshi education system and grade levels are entirely different from the U.S., the mapping is faulty and led to incorrect results. In contrast, we map grade level to different age groups to present the age-to-age comparison for the readability formulas to eliminate the incompatibility.

**Traditional Machine Learning Based Approaches** Sinha and Basu (2016) also used machine learning methods for Bengali readability classification, which are Support Vector Machine (SVM) and Support Vector Regression (SVR). They showed that features like average word length, number of consonant conjuncts play a significant role in Bengali readability analysis. Islam, Rahman, and Mehler (2014) used a combination of 18 lexical features and information-theoretic features to achieve a better score. Phani, Lahiri, and Biswas (2014) introduced the importance of inter-rater agreement study in the field of readability analysis of Bengali text. For agreement study, they used Cohen's kappa and Spearman's rank correlation coefficient. Recently, Phani, Lahiri, and Biswas (2019) proposed 11 readability measuring models based on regression analysis using 30 Bengali passages. They used features such as the presence of stop words, word sense, and POS tags. These prior works highlighted the importance of consonant conjunct for measuring readability. But they did not present any specific algorithm to compute consonant conjunct. Instead, in this paper, we present an efficient algorithm and human-annotated corpus to evaluate the effectiveness of the proposed algorithm. Another limitation of the works mentioned above is that their dataset is small scale and not publicly available. On the other hand, we present a large-scale dataset and design supervised neural models for Bengali readability prediction.

# Dataset

We collect documents from several published textbooks and magazines from Bangladesh and India. These are the most common and very well-known among children and adults. These documents usually were published after rigorous review and editorial process and widely read by various age groups. In this paper, we present two datasets for readability prediction. (1) Document-level dataset to experiment with formula-based approaches, (2) Sentence level dataset to train supervised neural models.

| Dataset | #Docs | Avg. #sents | Avg. #words |
|---|---|---|---|
| **NCTB** | 380 | 66.8 | 585.8 |
| **Additional** | 238 | 391.2 | 3045.0 |

Table 1: Statistics of the document-level dataset.

## Document-level Dataset

**NCTB** We select 16 textbooks from class 1 to 12 provided by *National Curriculum and Textbook Board (NCTB), Bangladesh*.[8] These textbooks are written by professional writers of **NCTB** who is responsible for the development of curriculum and distribution of textbooks. A majority of the Bangladeshi schools follow these books.[8] From class 3, 4, and 5, we take Bengali literature, Social Science, and General Science books; and from the rest of the classes, we take only Bengali literature books.

**Additional Sources** We also collect documents (literature and articles) for both children and adults from various well known and popular sources, some of which are: *Ichchhamoti*[9], *Society for Natural language Technology Research*[10], *Ebela*[11], *Sananda*[12], and *Bangla library*[13].

## Sentence-level Dataset

As we can see from Table 1, the document-level dataset is largely insufficient for training supervised neural models. Therefore, we further divide the documents into sentences to create a large-scale dataset for training neural models.

- **Simple Documents:** From the NCTB dataset, we select class 1 to 5 as simple documents as these documents are generally for 6 to 10 years old students.[7] Also, we take all the children type documents from the additional sources.
- **Complex Documents:** All adults type documents from additional sources, and we do not take any complex documents from the NCTB dataset.

Sentences from our simple documents are labeled as *'simple'*, and sentences from complex documents are labeled as *'complex'*. So initially, we start with 40,917 simple sentences and 60,875 complex sentences. After carefully investigating the initial dataset, we found that the complex documents mostly contain complex sentences. However, some simple sentences also exist in complex documents and vice versa. To remove simple sentences from the complex set, we apply

[8] https://w.wiki/ZwJ
[9] https://www.ichchhamoti.in/
[10] https://nltr.org/
[11] https://ebela.in/
[12] https://www.sananda.in/
[13] https://www.ebanglalibrary.com/

| | Train | Dev | Test |
|---|---|---|---|
| **Simple Sentences** | | | |
| #Sents | 37,902 | 1,100 | 1,100 |
| Avg. #words | 8.16 | 7.97 | 8.31 |
| Avg. #chars | 44.71 | 43.85 | 45.57 |
| **Complex Sentences** | | | |
| #Sents | 54,033 | 1,100 | 1,100 |
| Avg. #words | 8.04 | 8.08 | 8.16 |
| Avg. #chars | 44.01 | 44.65 | 44.63 |

Table 2: Statistics of the sentence-level dataset.

cosine similarity to every complex set sentences to every simple set sentences. We extract the sentences from the complex set, which has a semantic similarity score of 0.90 or more to the simple sentences. We manually recheck and annotate these extracted sentences to either simple or complex. Before measuring cosine similarity, we convert sentences to 300-dimensional vectors using a fastText pre-trained model for the Bengali language (Grave et al. 2018). It's important to note that these sentences we extracted are editor-verified and further annotated by us. Finally, after removing duplicate sentences, some simple sentences from complex sentences, and vise versa, we have 40,102 simple sentences and 56,233 complex sentences for training. While annotating, we also corrected the spelling mistakes to make the data clean. Table 2 indicates the summary of our sentence level dataset with train, dev, and test splits.

## Experiments

We use our document level dataset to experiment with formula-based approaches and use the sentence-level dataset to train supervised neural models.

### Formula-based Approaches

We select 10 documents (class 1 to 8, class 9/10, and class 11/12 in Table 3) from NCTB, and 4 documents (children 1 to adults 2 in Table 3) from additional sources. Due to the unavailability of the spoken syllable counting system for the Bengali language, we take a subset of the documents covering each class from the document level dataset. Flesch Reading Ease, Flesch–Kincaid Grade level, Gunning Fog Index, and SMOG grade formula require a common feature, which is the number of syllables. Counting syllables manually of all words from vocabulary is time-consuming. Google has provided NLP resources for various languages (e.g., Hindi, Urdu, Nepali, Sinhala, and Bengali). We use a pronunciation dictionary[14] for the Bengali language with 65,038 words. Although we use this dictionary, we have to manually count syllables for more than two thousand words, which are not present in that dictionary. We use the updated dictionary to experiment with formula-based approaches. Table 3 indicates the performance of formula-based approaches on our dataset. Here, we present a column **BN age**, which indicates the reader's age of the input documents. In Bangladesh, usually, children are admitted to Class 1 at the age of six, and complete their higher

**Algorithm 1:** Consonant Conjunct Count Algorithm.

```
Procedure ConsonantconjunctCount(W)
    Data: Input word W, which is an array of Bengali
          characters.
    Result: Return the number of consonant conjuncts in
            input word W.
    A ← Bengali sign VIRAMA;
    cc_count ← 0;
    l ← length(W);
    for k ← 0 to l − 1 do
        if W[k] == A then
            if k − 1 ≥ 0 and k + 1 < l then
                if k − 2 ≥ 0 then
                    if W[k − 1] and W[k + 1] is a
                     Bengali Consonant and W[k − 2]
                     != A then
                        cc_count ← cc_count + 1;
                    end
                end
                else if W[k − 1] and W[k + 1] is a
                 Bengali Consonant then
                    cc_count ← cc_count + 1;
                end
            end
        end
    end
    return cc_count;
```

secondary education (Class 12) at the age of seventeen.[7] Therefore, we fill up the age of the first 10 documents in Table 3 according to this range. For **Children 1** and **Children 2**, we set the age range 6-10 as children's literatures are created for 6 to 10 years old readers[15], and 18 or more for **Adults 1** and **Adults 2** in Table 3. Instead of matching the U.S. and Bangladeshi grade level, we match *'U.S. age'* and *'BN age'* for measuring the performance of all the formulas.

**Automated Readability Index (ARI)** formula provides a score equivalent to a grade level which was developed to assess the readability of written materials used in the U.S. Air Force (Senter and Smith 1967). The formula uses long words and long sentences to calculate a readability score. From Table 3, age range 6-7 indicates first or second grade.[16] From the second column of Table 3, the first row where the input document is taken from Class 1, so we set the value of 'BN age' = 6. After applying ARI to the input document, we find ARI score = 1 ('ARI' column). Since ARI score of 1 indicates Kindergarten grade level, therefore, we set 'U.S. age' equals to "5-6". As we can see, this document is correctly classified because 6 is present in the range of 5-6.

**Flesch Reading Ease (FE)** Formula approximates a number indicating the difficulty of the input document. The higher the number, the easier it is to read the document (Flesch 1948). For example, if the score is between 90 to 100 for an input text, then this text is very easy to read and understandable by an average 11 years old reader. The formula focuses on the

[14]https://git.io/JJhdm

[15]https://w.wiki/Z7W

[16]https://w.wiki/aRc

| Document | BN age | ARI | U.S. age | FE | U.S. age | FK | U.S. age | GF | U.S. age | SM OG | U.S. age | DC | U.S. age |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class 1 | 6 | 1 | **5-6** | 40.9 | 18-22 | 9 | 14-15 | 6 | 11-12 | N/A | - | 5.9 | 10-12 |
| Class 2 | 7 | 1 | 5-6 | 30.6 | 18-22 | 10 | 15-16 | 10 | 15-16 | 9 | 14-15 | 5.3 | 10-12 |
| Class 3 | 8 | 3 | **7-9** | 21.9 | ≥21 | 12 | 17-18 | 11 | 16-17 | 10 | 15-16 | 7.2 | 14-16 |
| Class 4 | 9 | 3 | **7-9** | 34.1 | 18-22 | 10 | 15-16 | 9 | 14-15 | 9 | 14-15 | 7.3 | 14-16 |
| Class 5 | 10 | 6 | 11-12 | 11.0 | ≥21 | 13 | 18-19 | 15 | 20-21 | 12 | 17-18 | 7.4 | 14-16 |
| Class 6 | 11 | 4 | 9-10 | 21.1 | ≥21 | 12 | 17-18 | 14 | 19-20 | 11 | 16-17 | 8.2 | 16-18 |
| Class 7 | 12 | 6 | **11-12** | 13.1 | ≥21 | 13 | 18-19 | 13 | 18-19 | 11 | 16-17 | 7.2 | 14-16 |
| Class 8 | 13 | 6 | 11-12 | 16.2 | ≥21 | 13 | 18-19 | 13 | 18-19 | 12 | 17-18 | 8.5 | 16-18 |
| Class 9/10 | 14-15 | 12 | 17-18 | -8.6 | - | 18 | ≥20 | 20 | ≥21 | 17 | ≥19-20 | 7.3 | **14-16** |
| Class 11/12 | 16-17 | 11 | **16-17** | -2.6 | - | 18 | ≥20 | 19 | ≥21 | 16 | ≥19-20 | 8.1 | **16-18** |
| Children 1 | 6-10 | 1 | **5-6** | 32.0 | 18-22 | 10 | 15-16 | 8 | 13-14 | 8 | 13-14 | 5.0 | **10-12** |
| Children 2 | 6-10 | 2 | **6-7** | 33.8 | 18-22 | 10 | 15-16 | 9 | 14-15 | 9 | 14-15 | 6.1 | 12-14 |
| Adults 1 | ≥18 | 12 | **17-18** | -22.8 | - | 21 | **≥20** | 24 | **≥21** | 19 | **≥19-20** | 11.5 | **≥21** |
| Adults 2 | ≥18 | 3 | 7-9 | 27.3 | **≥21** | 11 | 16-17 | 10 | 15-16 | 9 | 14-15 | 7.1 | 14-16 |

Table 3: Performance of the formula-based approaches. The bold-faced values indicate the correctly predicted U.S. age-groups for different formulas with Bengali (BN) age. The ARI formula performed reasonably well compared to other formulas.

number of words, sentences, and syllables. In our case, we get the maximum value 40.92 for an input document from Bangladeshi Class 1 (6 years old). According to this formula, a score of 40 means the text is difficult to read and is understandable by U.S. college students (18 to 24 years old). The lowest value of this formula is 0, but we get negative values for 3 documents out of 14 documents. This formula is highly suited for the documents written in English language and used by the professional readability analysis tools like Grammarly[2] and Readable[3]. However, the formula is not suitable for the Bengali language because of the various linguistic difficulty present in the syllable counting system, and this is one of the main reasons we experiment with wide varieties of formulas.

**Flesch–Kincaid (FK), Gunning Fog (GF), and SMOG** For all 3 formulas, the lower the number, the easier it is to read the document. For Flesch-Kincaid (Kincaid 1975), we get around 8.80 as our lowest value for Bangladeshi Class 1 level input document (6 years old reader). However, according to this rule, the input document having a score of 8 is for 8th U.S. grade level students (8th grade = 13/14 years old). Gunning fog index (Kincaid 1975) depends on the number of complex words, where a complex word means it has 3 or more syllables, and it is not a proper noun, compound word, or familiar jargon. We only consider 'syllable' and 'proper noun' for counting complex words. To identify proper nouns, we use a POS tagger provided by BNLP library.[17] We can not calculate SMOG (Mc Laughlin 1969) for the first document (Class 1) in Table 3, because this document has 28 sentences as SMOG formula requires at least 30 sentences.[18]

**Dale–Chall Formula (DC)** As mentioned before, the Dale–Chall formula (Dale and Chall 1948; Chall and Dale 1995) requires a 3,000 English words list which is familiar to U.S 4th grade (9-10 years old) students. As an alternative to this, we manually annotate 3,396 Bengali words. According to the Dale-Chall formula, any word that is not in our Bengali 3,396 words list is treated as a difficult word.

Table 3 indicates poor performance for Flesch reading ease, Flesch-Kincaid grade level, Gunning fog, and SMOG, therefore we can say that these formulas are not ideal for measuring the readability of the Bengali texts. On the other hand, the ARI performs relatively well as it correctly measures the age group of 8 documents out of 14 documents.

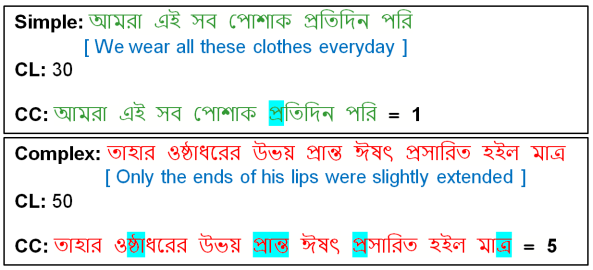


Figure 2: Visual representation of CL and CC for a Simple and a Complex Sentence.

## Supervised Neural Approaches

Due to the long-range dependencies of RNNs and the unavailability of large-scale human-annotated corpora, encoding of documents still lacks satisfactory solutions (Trinh et al. 2018). On the other hand, some document-level readability formulas, such as the SMOG index, require at least 30 sentences to measure a score.[18] In this section, we tackle these issues by dividing the document-level task into a supervised binary sentence classification problem where we have two classes, simple and complex. We design neural architectures to experiment with our sentence level dataset, which is presented in the dataset section.

**Additional Feature Fusion** The words present in complex sentences can be longer in terms of characters than the words in simple sentences. Therefore, we choose Character Length (**CL**), which indicates the total number of characters in a

[17]https://pypi.org/project/bnlp-toolkit/
[18]https://w.wiki/aRd

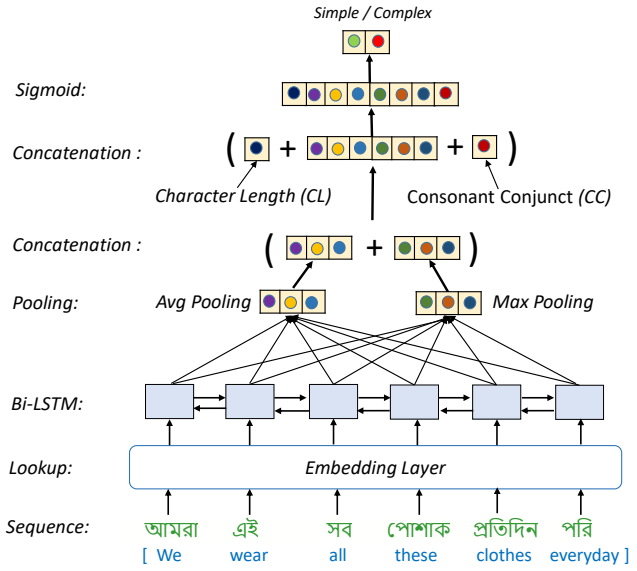


Figure 3: Readability prediction model.

| Baseline Models | | | | |
|---|---|---|---|---|
| **Models** | **A** | **R** | **P** | **F1** |
| BiLSTM | 77.5 | 69.4 | 82.8 | 75.5 |
| BiLSTM + Attention | 76.4 | 65.9 | 83.3 | 73.6 |
| **Ablations** | | | | |
| **Models** | **A** | **R** | **P** | **F1** |
| BiLSTM with Pooling | 81.3 | 78.8 | 83.0 | 80.8 |
| + Word2vec | 85.5 | 80.2 | 89.7 | 84.7 |
| + CL + CC | 85.7 | 80.9 | 89.5 | 85.0 |
| + GloVe | 86.1 | 79.3 | **91.9** | 85.1 |
| + CL + CC | 86.1 | 81.3 | 89.9 | 85.4 |
| + fastText | 86.0 | 80.1 | 90.8 | 85.1 |
| + CL + CC | **86.4** | 82.9 | 89.1 | 85.9 |
| + BPEmb | 86.2 | 81.5 | 90.0 | 85.6 |
| + CL + CC | 86.0 | 81.2 | 89.8 | 85.3 |
| + ULMFiT | 85.5 | 77.6 | **92.0** | 84.2 |
| + CL + CC | 86.2 | 80.4 | 91.0 | 85.4 |
| + TransformerXL | 86.3 | 82.7 | 89.0 | 85.8 |
| + CL + CC | **86.7** | 83.5 | 89.3 | **86.3** |
| + LASER | **86.4** | 84.3 | 88.0 | 86.1 |
| + CL + CC | 86.3 | **84.6** | 87.6 | 86.1 |
| + LaBSE | 86.0 | 80.3 | 90.8 | 85.2 |
| + CL + CC | **86.7** | **86.5** | 86.8 | **86.7** |

Table 4: Performance of Baseline and our ablations. The best results are marked green and second best results are marked blue. A, R, P, and F1 denote Accuracy, Recall, Precision, and F1 score respectively.

sentence, including white spaces as our additional readability feature. Moreover, the number of Consonant Conjunct (**CC**) is also an indicator for the complex sentence (Sinha and Basu 2016; Phani, Lahiri, and Biswas 2019). Unfortunately, these prior works did not present any specific algorithm to compute consonant conjunct for Bengali texts. Hence, we present a detailed procedure for counting consonant conjunct in Algorithm 1. To evaluate our consonant conjunct counting algorithm, we manually create a dataset with 341 words[19] and their corresponding consonant conjunct count. Our algorithm can successfully count the CCs present in all 341 words from our dataset. Figure 2 presents a visual representation of the counting of CC and CL for two example sentences.

**Ablation Experiments** We select Bidirectional LSTM (BiLSTM) (Schuster and Paliwal 1997) and BiLSTM with attention mechanism (Raffel and Ellis 2016) as baseline models. As we can see from Table 4, BiLSTM model has achieved better performance from the baseline models, therefore, we extend the BiLSTM model by adding global average pooling and global max-pooling layers (Boureau, Ponce, and LeCun 2010). We use this model for our ablation experiments with the additional features to demonstrate the effects of CL and CC feature fusion. For each input sentence, we calculate the CL and CC to concatenate with the pooling layers. Given an input sentence $s_j$, its word sequence $w_1, w_2, w_3 ... w_{|s_j|}$ is fed into a word embedding layer to obtain embedding vectors $x_1, x_2, x_3 ... x_{|s_j|}$ before passing it to the BiLSTM layer. The word embedding layer is parameterized by an $\mathbf{E}_w \in \mathbb{R}^{K \times |V|}$ embedding matrix, where K is the embedding dimension and $|V|$ is the vocabulary size. The overall process is summarized in Figure 3. We use all pre-trained language models available to date for the Bengali language, which includes:

- 300 dimensional Word2vec pre-trained model (Mikolov et al. 2013).
- 300 dimensional GloVe pre-trained model (Pennington, Socher, and Manning 2014).
- 300 dimensional vector representation from fastText pre-trained model (Grave et al. 2018).
- 300 dimensional BPEmb (Heinzerling and Strube 2018) model, which is based on Byte-Pair encoding, provides a collection of pre-trained subword embedding models for 275 languages including Bengali.
- 400 dimensional pre-trained ULMFiT (Howard and Ruder 2018) model provided by iNLTK[20].
- 410 dimensional pre-trained TransformerXL (Dai et al. 2019) model provided by iNLTK[20] library.
- 1024 dimensional Language-agnostic Sentence Embedding model laserembeddings[21] for 93 languages, which is based on LASER (Artetxe and Schwenk 2019).
- 768 dimensional Language-agnostic BERT Sentence Embedding model LaBSE for 109 languages (Feng et al. 2020).

**Hyperparameters** We use 60 as maximum sequence length with a batch size of 16, embedding size of 300, 64 LSTM hidden units, and Adam optimizer (Kingma and Ba 2015) with a learning rate of 0.001. We run the training for 50 epochs and check the improvement of validation (dev set)

[19] https://w.wiki/Wk8

[20] https://git.io/JUItc

[21] https://pypi.org/project/laserembeddings/

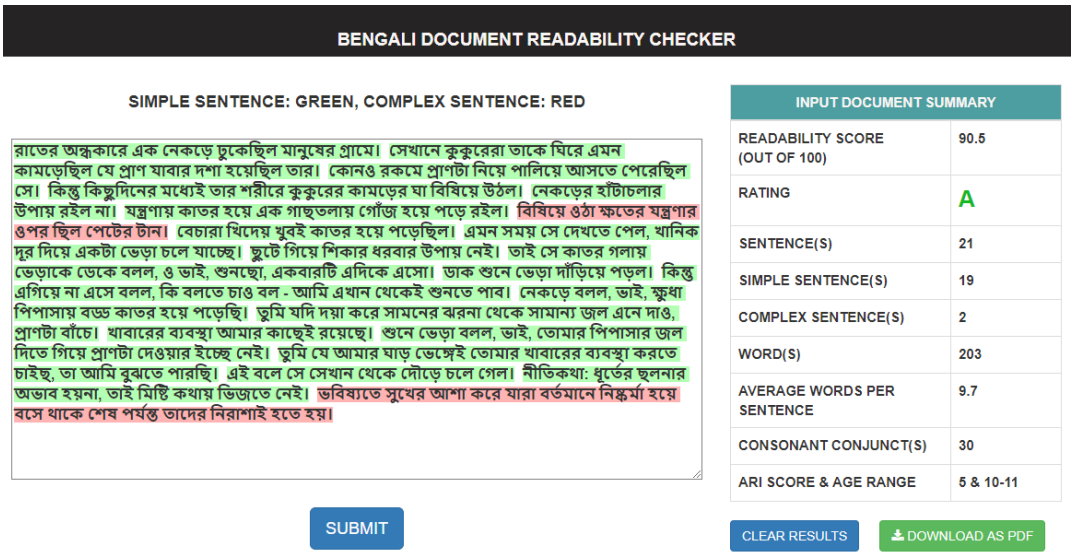


Figure 4: Bengali Document Readability Checker. For an input document, the simple sentences are highlighted with green color and the complex sentences are highlighted with red color. The document will be easy to read for people with age between 10-11.

| Models | A | R | P | F1 |
|---|---|---|---|---|
| fastText Unigram | 86.0 | 82.8 | 88.4 | 85.5 |
| fastText Bigram | 86.6 | 84.9 | 87.9 | 86.4 |
| fastText Trigram | **87.4** | **85.0** | **89.2** | **87.1** |

Table 5: Performance of Supervised Pretraining.

loss to save the latest best model during training. It is important to note that we use the same hyperparameters for the baseline models.

**Results** We present the detailed ablation experiment results of our test set in Table 4. We didn't perform any explicit prepossessing before computing the semantic representation using the embedding layer lookup in Figure 3. We compute accuracy, precision, recall, and F1 score to compare the performance of our readability prediction model with the baselines. We perform a detailed ablation study with the variations of pre-trained embedding models and our additional feature fusions. After applying all the models from Table 4 to our test set, we get maximum accuracy (86.7%) and F1 score (86.7%) from the combination of BiLSTM with pooling, CL, CC, and embeddings from LaBSE model. As shown in Table 4, the impact of additional features such as CL and CC of our readability prediction model is significant, achieving maximum Accuracy and F1 scores 6 times out of 8 cases.

**Supervised Pretraining** We also experiment with fastText supervised text classification techniques (Joulin et al. 2017). These models are based on multinomial logistic regression. We build 3 classifiers using word n-grams (unigram, bigram, and trigram) and character n-grams (2 to 6 length) with learning rate = 0.5 and 50 epochs. For all 3 cases, we use hierarchical softmax (Peng et al. 2017) as the loss function for faster training. From Table 5, we observe that the model with word trigram and character n-grams achieves maximum Accuracy and F1 score.

## Bengali Readability Analysis Tool

We design a Bengali readability analysis tool capable of providing in-depth information of readability and complexity of the documents which is shown in Figure 4. For an input document $D$ with $N$ sentences, our Bengali readability checker can highlight the simple and complex sentences. We calculate the readability score based on the ratio of the predicted simple sentences to the total number of sentences in the document $D$ and assign different ratings based on the readability scores. Since ARI performs well on the Bengali document-level readability analysis and it is resource independent, therefore, we use the scores returned from this formula to predict the age-group of the input document. The example document $D$ in Figure 4 will be reasonably easy to read for most people at least with age between 10-11.

## Conclusion

In this paper, we present a readability analysis tool that would be useful for educators, content writers or editors, researchers, and readers of different ages. We adopt document-level readability formulas traditionally used for U.S. based education system to the Bengali education system with a proper age-to-age comparison. Moreover, we divide the task into sentence-level and design supervised neural models, which will serve as a baseline for the future works of this task. We present several human-annotated corpora, dictionaries, and an algorithm, which can be useful for several other tasks of this low-resource language. In the future, we wish to improve the quality of our system by increasing the size of our sentence-level dataset and will present a user-study based on our tool. Also, we will focus on the readability analysis of Bengali-English code-mixed texts.

## Acknowledgments

We would like to thank all the anonymous reviewers for their thoughtful comments and constructive suggestions.

## References

Al-Twairesh, N.; Al-Dayel, A.; Al-Khalifa, H.; Al-Yahya, M.; Alageel, S.; Abanmy, N.; and Al-Shenaifi, N. 2016. MADAD: A Readability Annotation Tool for Arabic Text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4093–4097. Portorož, Slovenia: European Language Resources Association (ELRA).

Artetxe, M.; and Schwenk, H. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* 7: 597–610.

Battisti, A.; Pfütze, D.; Säuberli, A.; Kostrzewa, M.; and Ebling, S. 2020. A Corpus for Automatic Readability Assessment and Text Simplification of German. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 3302–3311. Marseille, France: European Language Resources Association.

Boureau, Y.-L.; Ponce, J.; and LeCun, Y. 2010. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 111–118.

Broda, B.; Nitoń, B.; Gruszczyński, W.; and Ogrodniczuk, M. 2014. Measuring Readability of Polish Texts: Baseline Experiments. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 573–580. Reykjavik, Iceland: European Language Resources Association (ELRA).

Chall, J. S.; and Dale, E. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Chowdhury, R. R.; Nayeem, M. T.; Mim, T. T.; Chowdhury, M. S. R.; and Jannat, T. 2021. Unsupervised Abstractive Summarization of Bengali Text Documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Online: Association for Computational Linguistics.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; and Salakhutdinov, R. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2978–2988. Florence, Italy: Association for Computational Linguistics.

Dale, E.; and Chall, J. S. 1948. A formula for predicting readability: Instructions. *Educational research bulletin* 37–54.

Das, S.; and Roychoudhury, R. 2006. Readability modelling and comparison of one and two parametric fit: A case study in Bangla. *Journal of Quantitative Linguistics* 13(01): 17–34.

Dueppen, A. J.; Bellon-Harn, M. L.; Radhakrishnan, N.; and Manchaiah, V. 2019. Quality and readability of English-language Internet information for voice disorders. *Journal of Voice* 33(3): 290–296.

El-Haj, M.; and Rayson, P. 2016. OSMAN—A Novel Arabic Readability Metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 250–255.

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2020. Language-agnostic BERT Sentence Embedding. *arXiv preprint arXiv:2007.01852* .

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology* 32(3): 221.

Fourney, A.; Ringel Morris, M.; Ali, A.; and Vonessen, L. 2018. Assessing the Readability of Web Search Results for Searchers with Dyslexia. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, 1069–1072. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356572.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Grigonyte, G.; Kvist, M.; Velupillai, S.; and Wirén, M. 2014. Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 74–83. Gothenburg, Sweden: Association for Computational Linguistics.

Heinzerling, B.; and Strube, M. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Howard, J.; and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. Melbourne, Australia: Association for Computational Linguistics.

Islam, Z.; Mehler, A.; and Rahman, R. 2012. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 545–553.

Islam, Z.; Rahman, M. R.; and Mehler, A. 2014. Readability classification of bangla texts. In *International conference on intelligent text processing and computational linguistics*, 507–518. Springer.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Valencia, Spain: Association for Computational Linguistics.

Kincaid, J. 1975. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch

report. Chief of Naval Technical Training, Naval Air Station Memphis.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* .

Loughran, T.; and McDonald, B. 2014. Measuring Readability in Financial Disclosures. *The Journal of Finance* 69(4): 1643–1671. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/43611199.

Mc Laughlin, G. H. 1969. SMOG grading-a new readability formula. *Journal of reading* 12(8): 639–646.

Meng, C.; Chen, M.; Mao, J.; and Neville, J. 2020. ReadNet: A Hierarchical Transformer Framework for Web Article Readability Analysis. In *European Conference on Information Retrieval*, 33–49. Springer.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 3111–3119. Red Hook, NY, USA: Curran Associates Inc.

Napolitano, D.; Sheehan, K.; and Mundkowsky, R. 2015. Online Readability and Text Complexity Analysis with TextEvaluator. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 96–100. Denver, Colorado: Association for Computational Linguistics.

Nguyen, P.; and Uitdenbogerd, A. L. 2019. Measuring English Readability for Vietnamese Speakers. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, 136–145.

Okinina, N.; Frey, J.-C.; and Weiss, Z. 2020. CTAP for Italian: Integrating Components for the Analysis of Italian into a Multilingual Linguistic Complexity Analysis Tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 7123–7131. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Peng, H.; Li, J.; Song, Y.; and Liu, Y. 2017. Incrementally Learning the Hierarchical Softmax Function for Neural Language Models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 3267–3273. AAAI Press.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Phani, S.; Lahiri, S.; and Biswas, A. 2014. Inter-rater Agreement Study on Readability Assessment in Bengali. *arXiv preprint arXiv:1407.1976* .

Phani, S.; Lahiri, S.; and Biswas, A. 2019. Readability Analysis of Bengali Literary Texts. *Journal of Quantitative Linguistics* 26(4): 287–305.

Raffel, C.; and Ellis, D. P. 2016. Feed-forward networks with attention can solve some long-term memory problems. *Workshop track - ICLR 2016* .

Rets, I.; Coughlan, T.; Stickler, U.; and Astruc, L. 2020. Accessibility of Open Educational Resources: how well are they suited for English learners? *Open Learning: The Journal of Open, Distance and e-Learning* 0(0): 1–20.

Reynolds, R. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 289–300. San Diego, CA: Association for Computational Linguistics.

Riddle, D. R. 2007. *Brain aging: models, methods, and mechanisms*. CRC Press.

Sato, S. 2014. Text Readability and Word Distribution in Japanese. In Calzolari, N.; Choukri, K.; Declerck, T.; Loftsson, H.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, 2811–2815. European Language Resources Association (ELRA).

Sato, S.; Matsuyoshi, S.; and Kondoh, Y. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45(11): 2673–2681.

Senter, R.; and Smith, E. A. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Seretan, V. 2012. Acquisition of Syntactic Simplification Rules for French. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Sinha, M.; and Basu, A. 2016. A study of readability of texts in Bangla through machine learning approaches. *Education and information technologies* 21(5): 1071–1094.

Sinha, M.; Sharma, S.; Dasgupta, T.; and Basu, A. 2012. New Readability Measures for Bangla and Hindi Texts. In *Proceedings of COLING 2012: Posters*, 1141–1150. Mumbai, India: The COLING 2012 Organizing Committee.

Trinh, T.; Dai, A.; Luong, T.; and Le, Q. 2018. Learning Longer-term Dependencies in RNNs with Auxiliary Losses. In *International Conference on Machine Learning*, 4965–4974.

Vajjala, S.; and Lučić, I. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 297–304. New Orleans, Louisiana: Association for Computational Linguistics.